Enhancing Patent Retrieval by Citation Analysis

Atsushi Fujii
Graduate School of Library, Information and Media Studies
University of Tsukuba
1-2 Kasuga, Tsukuba, 305-8550, Japan
fujii@slis.tsukuba.ac.jp

ABSTRACT

This paper proposes a method to combine text-based and citation-based retrieval methods in the invalidity patent search. Using the NTCIR-6 test collection including eight years of USPTO patents, we show the effectiveness of our method experimentally.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models

General Terms

Experimentation, Measurement

Keywords

Patent retrieval, Citation analysis, NTCIR

1. INTRODUCTION

Processes of patent retrieval differ significantly, depending on the purpose of retrieval. One process is the "technology survey", in which patents related to a specific technology, e.g., "blue light-emitting diode", are searched. This process is similar to ad hoc retrieval tasks targeting nonpatent documents.

Another process is the "invalidity search", in which prior arts related to a patent application are searched. Away from academic research, invalidity searches are performed by examiners in government patent offices and searchers in the intellectual property divisions of private companies.

The use of patents in information retrieval research dates back at least to the 1970s [1], when 76 US patents were used to evaluate the effectiveness of local feedback techniques. The experiment in this research simulated the technology survey task, in which seven queries, such as "electrode structure", were used. For each query, an electronics engineer selected relevant patents from the 76 US patents.

Osborn and Strzalkowski [5] performed experiments on the invalidity search task. They used approximately 6000 US patents as a target document collection. Each search topic was also a patent. The patents cited in the topic patent (i.e., citations) were used as relevant documents, because citations are usually prior arts for a citing patent. Thus, no relevance judgment by human experts was needed.

In the Sixth NTCIR Workshop (NTCIR-6), the Patent Retrieval Task was organized and three subtasks were performed; Japanese Retrieval, English Retrieval, and Classification [3][1]. The English Retrieval subtask intended the invalidity search targeting US patents. However, the number of target documents was larger than those used in previous experiments.

In this paper, we propose our retrieval method participated in NTCIR-6. Our method uses both text content and citations to enhance the invalidity search. We also show the effectiveness of our method experimentally.

2. NTCIR-6 PATENT RETRIEVAL TASK

In the NTCIR-6 English Retrieval subtask, target documents are USPTO patents published in 1993–2000. The number of documents is 981,948. In each document, a number of additional SGML-style tags are inserted to specify the fields, such as bibliographic information and text content.

Each search topic is one or more claims extracted from a patent published in 2000–2001. The organizers of the Patent Retrieval Task used a number of criteria to select 2221 patents as search topics.

The pooling-based relevance judgement was not performed and relevant documents for a search topic are the citations in the search topic. If a topic patent and its relevant document are assigned to the same IPC (International Patent Classification) code, the document can usually be retrieved with a high accuracy. Thus, the degree of the relevance of each citation is classified into the following ranks.

- A: The IPC subclasses assigned to the topic patent and the target document are not identical.
- B: The IPC subclasses assigned to the topic patent and the target document are identical.

The primary IPC code for each patent is identified in the bibliography field.

3. METHODOLOGY

Traditional research in citation analysis can be used in different applications for patents [4]. For example, if a patent is cited by a large number of other patents, this cited patent is possibly a foundation of those citing patents and is, therefore, important.

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.
ACM 978-1-59593-597-7/07/0007.

[1] http://if-lab.slis.tsukuba.ac.jp/fujii/ntc6pat/cfp-en.html

This idea is similar to identifying authoritative pages by analyzing hyperlink structures on the World Wide Web. For example, Yang [7] combined text-based and link-based methods in the Web retrieval.

Following the above ideas, we combine text and citation information in the invalidity patent search.

For the text-based retrieval, we use the claim(s) in each document to perform word-based indexing. We use Okapi BM25 [6] to compute the text-based score for each document with respect to a query.

For the citation-based retrieval, we use two alternative methods. In either method, we first perform the text-based retrieval and obtain top N documents. We then compute the citation-based score for each of the N documents. Finally, we combine the text-based and citation-based scores and resort the N documents. We compute the final score for document d, $S(d)$, by Equation (1).

$$S(d) = S_T(d) \times S_C(d)^{\alpha} \quad (1)$$

$S_T(d)$ and $S_C(d)$ denote the text-based and citation-based scores for d, respectively. α is a parametric constant to control the effects of S_C.

As a citation-based method, we use PageRank [2], which estimates the probability that a user surfing on the Web visits a document. We use this probability as the citation-based score for each document. Given a document collection, the value of PageRank for each document is a constant and is independent of the topic.

As an alternative citation-based method, we propose a topic-sensitive method. We use only citations among the top N documents. As in PageRank, the citation-based score of document d is determined by the total votes by other documents. If d is cited by a large number of documents, a high score is given to d. However, if a document cites n documents, the vote for each cited document is $\frac{1}{n}$. We compute $S_C(d)$ by Equation (2).

$$S_C(d) = \sum_{x \in D_{*\rightarrow d}} \frac{1}{|D_{x\rightarrow *}|} \quad (2)$$

$D_{*\rightarrow d}$ and $D_{d\rightarrow *}$ denote a set of documents citing d and a set of documents cited by d, respectively.

4. EXPERIMENTS

Using the NTCIR-6 test collection described in Section 2, we compared the effectiveness of the following methods.

- (a) text-based retrieval
- (b) text-based retrieval + PageRank
- (c) text-based retrieval + topic-sensitive citation-based method

For all methods, $N = 1000$. We determined the optimal value of α in Equation (1) through preliminary experiments. The values of α were 0.01 and 0.1 for methods (b) and (c), respectively.

Tables 1 and 2 show different evaluation measures for the above three methods. While in Table 1 we used only A documents as correct answers, in Table 2 we used both A and B documents as correct answers.

In Tables 1 and 2, "R@X" denotes the recall at the top X documents. For the value of X, we used 100, 200, and 500, because examiners and searchers usually investigate a couple of hundreds of documents for a single topic. We also used Mean Average Precision (MAP) as an evaluation measure.

Table 1: Evaluation results for rigid relevance.

Method	R@100	R@200	R@500	MAP
(a)	0.1548	0.2108	0.2960	0.0340
(b)	0.1650	0.2236	0.3096	0.0355
(c)	0.1753	0.2336	0.3211	0.0389

Table 2: Evaluation results for relaxed relevance.

Method	R@100	R@200	R@500	MAP
(a)	0.1965	0.2634	0.3700	0.0712
(b)	0.2057	0.2770	0.3838	0.0748
(c)	0.2171	0.2910	0.3991	0.0811

Looking at Tables 1 and 2, methods (b) and (c) were more effective than method (a), irrespective of the evaluation measure and the relevance degree. However, method (c) was more effective than method (b). We used the paired t-test for statistical testing. Although the MAP values of (a) and (b) in Table 1 were not significantly different, for the other cases the difference was statistically significant at the 1% level.

5. CONCLUSION

We used eight years of USPTO patents and demonstrated the effectiveness of the citation analysis in the invalidity patent search. A combination of the text-based and citation-based methods improved the text-based method. The improvement was even greater when we used the topic-sensitive citation-based method.

6. REFERENCES

[1] R. Attar and A. S. Fraenkel. Local feedback in full-text retrieval systems. *Journal of the Association for Computing Machinery*, 24(3):397–417, 1977.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks*, 30(1–7):107–117, 1998.

[3] A. Fujii, M. Iwayama, and N. Kando. Overview of the patent retrieval task at the NTCIR-6 workshop. In *Proceedings of the Sixth NTCIR Workshop Meeting*, 2007.

[4] M. M. S. Karki. Patent citation analysis: A policy analysis tool. *World Patent Information*, pages 269–272, 1997.

[5] M. Osborn and T. Strzalkowski. Evaluating document retrieval in patent database: a preliminary report. In *Proceedings of the 6th ACM International Conference on Information and Knowledge Management*, pages 217–221, 1997.

[6] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proceedings of the 3rd Text REtrieval Conference*, 1994.

[7] K. Yang. Combining text- and link-based retrieval methods for Web IR. In *Proceedings of the 10th Text REtrieval Conference*, pages 609–618, 2001.